Advanced Engineering Informatics 30 (2016) 327–336

Contents lists available at ScienceDirect

Advanced Engineering Informatics

journal homepage: www.elsevier.com/locate/aei

Vision-based action recognition of construction workers using dense trajectories

Jun Yang [a,*], Zhongke Shi [b], Ziyan Wu [a]

[a] School of Mechanics, Civil Engineering and Architecture, Northwestern Polytechnical University, China
[b] School of Automation, Northwestern Polytechnical University, China

ARTICLE INFO

Article history:
Received 12 July 2015
Received in revised form 25 April 2016
Accepted 26 April 2016

Keywords:
Worker
Action recognition
Construction
Computer vision
Dense trajectories

ABSTRACT

Wide spread monitoring cameras on construction sites provide large amount of information for construction management. The emerging of computer vision and machine learning technologies enables automated recognition of construction activities from videos. As the executors of construction, the activities of construction workers have strong impact on productivity and progress. Compared to machine work, manual work is more subjective and may differ largely in operation flow and productivity among different individuals. Hence only a handful of work studies on vision based action recognition of construction workers. Lacking of publicly available datasets is one of the main reasons that currently hinder advancement. The paper studies worker actions comprehensively, abstracts 11 common types of actions from 5 kinds of trades and establishes a new real world video dataset with 1176 instances. For action recognition, a cutting-edge video description method, dense trajectories, has been applied. Support vector machines are integrated with a bag-of-features pipeline for action learning and classification. Performances on multiple types of descriptors (Histograms of Oriented Gradients – HOG, Histograms of Optical Flow – HOF, Motion Boundary Histogram – MBH) and their combination have been evaluated. Discussion on different parameter settings and comparison to the state-of-the-art method are provided. Experimental results show that the system with codebook size 500 and MBH descriptor has achieved an average accuracy of 59% for worker action recognition, outperforming the state-of-the-art result by 24%.

© 2016 Elsevier Ltd. All rights reserved.

1. Introduction

Productivity in the construction industry has been declining during the past few decades [1]. Since labor accounts for 33–50% of the total cost of a project, their productivity is a key factor in schedule and budget control [2]. One efficient way to manage workers' performance is to monitor their activity on site, analyze the operation in real time, optimize the work flow dynamically [3–6]. Historical observation can also benefit future worker training and education.

To monitor worker activities, current efforts usually lean on foremen collecting information from construction site by means of onsite observations, survey or interview [7]. Post processing is often required to analyze the collected data manually. The entire procedure is labor intensive, cost sensitive and can be prone to error. As reported in [8], for a case study of 870 m^2 tiling trade, 336 manual observations are required to measure the six workers' productivity. The observation has to be made four rounds a day, lasting for 14 days. Not to mention each observation has to record the specific task in detail, as well as the active and inactive time. There is an urgent need of automated activity analysis of construction workers.

Recent years, with the prevalence of cameras in construction sites, images and videos become low-cost and reliable information resources. The emerging of computer vision and machine learning technologies enables analyzing construction activities automatically. In the past decade, many researchers have dedicated to this field and made remarkable achievements [9–11]. However, some open challenges remain unsolved. For example, the behavior of construction workers needs to be further explored.

Recognition of worker behavior can be performed at various levels of abstraction. As suggested by Moeslund et al. [12], there are "action primitives", "actions", and "activities". An action primitive is an atomic movement usually in limb level, e.g., pick up a brick. An action is composed by a series of action primitives, either sequential different primitives or repetitive single primitive, e.g.,

* Corresponding author.
E-mail addresses: junyang@nwpu.edu.cn (J. Yang), zkshi@nwpu.edu.cn (Z. Shi), zywu@nwpu.edu.cn (Z. Wu).

http://dx.doi.org/10.1016/j.aei.2016.04.009
1474-0346/© 2016 Elsevier Ltd. All rights reserved.

laying a brick contains steps of "pick up a brick", "get mortar with a trowel", "smear the mortar", "place the brick", and "knock the brick with the trowel to fasten". An activity is in the highest level, involving in a number of subsequent actions, e.g., building a wall requires measuring, alignment, and laying bricks.

In this paper, we focus on worker action recognition from pre-segmented video clips. If integrating with action detection or segmentation in longer videos, worker productivity can be assessed automatically. Furthermore, action recognition can form initial steps towards worker activity analysis.

The contributions of this paper are twofold. First, a large scale dataset of worker actions covering a wide range of trades has been constructed. Existing human action data sets mainly focus on general body movements (walking, waving, turn around) [13–15] or common daily activities (sports [16,17], cooking [18–22], etc.). Datasets on specialty activities are rare, which by their nature have smaller inter-class difference and introduce difficulties in recognition. Second, how existing action recognition algorithm will perform on a large scale construction dataset is unknown, especially when both coarse-grained and fine-grained actions coexist. By adopting a cutting-edge video representation method – dense trajectories and evaluating on various feature descriptors, we achieve an average accuracy of 59% for worker action recognition, outperforming the state-of-the-art result by 24%.

The proposed worker action dataset is available upon request. A preliminary version of this article has appeared in [23].

The rest of the paper is organized as follows. Section 2 reviews the related literatures and discusses existing challenges. Section 3 describes the methodology in detail by illustrating dense trajectories algorithm and related feature descriptors, as well as the classification method. Section 4 presents the new data set. Section 5 gives out experimental results with discussion on parameters setting and comparison against state-of-art results. Section 6 concludes the paper.

2. Related work

This section introduces the state-of-the-art human action recognition from different aspects and discusses open challenges in worker action recognition.

2.1. Action recognition in computer vision field

Action recognition has gained plenty of interest in computer vision field due to its potential in a wide range of applications, such as robotics, video surveillance, and human–computer interface [24,25]. During the past decades, numerous approaches have been proposed for human action recognition. One of the most successful line of work is the Bag-of-Feature (BoF) [26], which detects local features in video frames, represents videos with feature descriptors, generates codebook by clustering on features and obtains a sparse histogram representation over the codebook for learning and classification. Action is spatial movement across time. Local spatio-temporal features encode video information at a given location in space and time [27]. Therefore they are suitable for action recognition. Feature detection approaches range from extended Harris detector [28], Gabor filter-based detector [29] to Hessian matrix based detector [30]. Some widely used feature descriptors are higher order derivatives, gradient information, optical flow and brightness [14,26,29]. Other researchers extend successful image descriptors to spatio-temporal domain for action recognition, such as 3D-SIFT [31], HOG3D [32], extended SURF [30], and Local Trinary Patterns [33]. Instead of representing features in the joint 3D space–time domain (wherein spatial information in images is 2D), a more intuitive option is to track feature points across time. Wang et al. [34] proposed to track the densely sampled feature points across the optical field and represent features combining multiple descriptors. Their method achieved a state-of-the-art performance on several common datasets. However, how it will score on specialty activities is still unknown.

2.2. Vision-based construction operation analysis

During the past decade, many researchers have applied computer vision technologies for construction operation analysis. For more comprehensive reviews, please refer to [9–11]. One main stream method is to detect, track workers and equipment and analyze their activities by poses or trajectories combining prior knowledge. Zou and Kim [35] track the excavator by appearance and judge the idle time through its movement status. Azar et al. [6] detect and track the excavator and dump truck simultaneously to analyze the dirt loading cycle. Gong and Caldas [36,37] detect a concrete bucket in video streams through machine learning and estimate its travel cycles based on the prior knowledge of construction site layout. Yang et al. [38] perform similar work of monitoring concrete placement activity by tracking the crane jib through 3D pose estimation. Peddi et al. [39] track workers tying rebar through blob matching, extract skeletons for pose estimation and classify their working status into effective, ineffective and contributory by poses. Gong and Calda [40] evaluate several popular algorithms for construction object recognition and tracking and develop a prototype system for construction operation analysis. Bugler et al. [41] propose a novel scheme to combine tracking based activity monitoring with photogrammetry based progress measurement for excavation process analysis.

However, in cluttered construction scenarios, it is difficult to detect and track construction entities stably through a long duration [42]. Errors from previous stages (detection and tracking) might accumulate and affect the activity analysis adversely. To solve this problem, a recent trend is to adopt the Bag-of-Feature pipeline for action recognition without detecting or tracking any construction entities explicitly. Gong et al. [43] utilize the 3D-Harris detector [28] as the feature detector, HoG (Histogram of Gradient) and HoF (Histogram of Optical Flow) as the feature descriptor, and Bayesian network models as the learning method for worker and backhoe action recognition. Golparvar-Fard et al. [44] focus on action recognition of earthmoving equipment. They use Gabor filter as feature detector [29], HoG and HoF as descriptor and Support Vector Machines for action learning. Both the above mentioned works [43,44] are tested on relatively small datasets. The average numbers of action types per each dataset are four and three separately. What is more, they all adopt a joint spatio-temporal feature description. The space domain and the time domain in videos have different characteristics naturally. It may not be reasonable to simply join them together.

Apart from obtaining videos by common cameras, adopting RGB-D cameras becomes a new trend in construction operation analysis [4,3,45,46]. Since RGB-D cameras can capture depth information, skeleton information is usually extracted to infer body poses related to various worker actions.

2.3. Datasets for action recognition

As a prerequisite for evaluation and comparison, a large amount of human action datasets have been created [47]. The complexity of existing datasets increases as that of the corresponding algorithms. Early age data sets concern more for full body actions and are usually captured under control environments. Typical examples are the Weizmann dataset [13], the KTH dataset [14] and the UIUC dataset [15]. Soon after there comes a need for real-world videos with less limitation on environment,

illumination, point of view and even performers (actors). Thanks to the prevalence of digital movies and web videos, many comprehensive datasets with diverse action categories have been collected, such as the YouTube dataset [48], the Holloywood2 dataset [49], the UCF sports dataset [16], the Olympic sports [17], the UCF50 dataset [50] and the HMDB51 dataset [51]. Surveillance datasets care more for the interaction between people [52,53] and the trajectories of movement, and are usually recorded from wider view angle and in longer duration [54–56].

Instead of focusing on general body actions, a recent trend is to explore goal-directed actions. Assisted daily living (ADL) datasets [57] fall into this category. Compared to coarse-grained activities for full-body motions, ADL activities are more fine-grained and share smaller inter-class variability, which propose serious challenges for action recognition. Kitchen duty is the most heavily studied scenario. Typical datasets include the CMU-MMAC dataset [18], the TUM kitchen dataset [19], the MPII dataset [20], the 50 salads dataset [21] and the Serre Breakfast dataset [22]. These datasets record human subjects preparing various types of food following given recipes. Beyond daily living activities, very few researchers study on professional activities from different fields, such as the automobile industry [58], the medical surgery [59] and the construction operations [43,44]. However, only the WR (Workflow Recognition) dataset [58] from the production line of the automobile manufacturer is made publicly available.

2.4. Open challenges for action recognition of construction workers

Except common challenges in action recognition, such as illumination change, various view angels and self-occlusion, several unique issues related with worker action recognition are discussed as follows:

- Worker actions are combined with both coarse-grained and fine-grained movements. For example, transporting materials mainly involve lifting and walking, which are coarse-grained. But tying rebars requires fine-grained arm and finger movements. This varies from previous studies, which focused either on coarse-grained actions [51,50] or fine-grained actions [21,22] only. Since coarse-grained and fine grained actions may require different granularity of feature description, it is interesting to explore how a single algorithm will perform on mixed data.
- The inter-class variability of different worker actions might be small while the intra-class variability might be big, which is extremely challenging for action recognition. For example, a worker bolting rebars may visually resemble another worker hammering nails since they all bent over with a tool in hand. While even performing the same task, the work flow from different individuals may vary largely due to personal habits.
- Worker actions vary from trades to trades. Previous work [43] only studies five action categories of formwork workers. There are many other types of trades to be explored, such as carpenter and ironworker. A comprehensive publicly available dataset is missing.

Aiming at the above mentioned challenges, the paper proposes a comprehensive worker action dataset with 1176 video clips, covering 11 types of worker actions from various trades. Both coarse-grained and fine-grained actions are involved. Meanwhile, multiple factors which may affect the action presentation in video are considered, such as illumination, view angle, workers' gender and skill level. To evaluate on existing action recognition algorithms, dense trajectories is adopted for video representation hence dense sampling exhibits superiority over sparse feature points and achieves state-of-the-art performance on multiple datasets [34].

3. Methodology

The overall workflow of the system is shown in Fig. 1. As it can been seen, the system is built upon a Bag-of-Feature structure. First, video clips are represented by visual feature descriptors. Specifically, dense trajectories are generated by dense sampling and tracking on a dense optical flow field. Then various descriptors are computed along the dense trajectories. Second, codebooks per each description channel are constructed using k-means clustering algorithm and descriptors are quantized by assigning to the nearest vocabulary word using Euclidean distance. Third, a non-linear SVM (Support Vector Machines) is adopted for classification. More details are illustrated as follows.

3.1. Dense trajectories generation

Unlike [43,44] which use spatial–temporal features for video representation, we adopt the dense trajectories method [34] to model videos. To make the paper self contained, a brief description is given below.

Dense trajectories are obtained by densely sampling and tracking feature points on multiple spatial scales separately. For every spatial scale, feature points are densely and equally sampled by step of W pixels. And the scale increases by a factor of $1/\sqrt{2}$. For future tracking, sampled points in homogeneous image areas is removed.

Dense optical field enables densely tracking features effortlessly, and ensures robust tracking of fast irregular motion patterns due to its smoothness nature. For an image frame I_t, suppose that its dense optical flow field $\omega_t = (u_t, v_t)$ is extracted, where u_t and v_t represent the horizontal and vertical component separately. For a given point $P_t = (x_t, y_t)$ in this frame, its tracked position in the next frame I_{t+1} is smoothed by median filtering on ω_t:

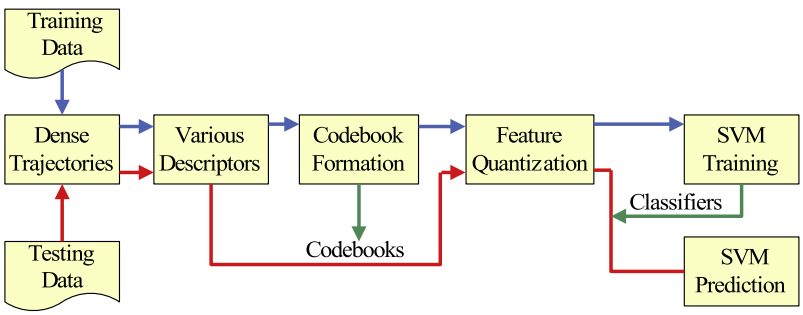

Fig. 1. The system overview.

$$P_{t+1} = (x_{t+1}, y_{t+1}) = (x_t, y_t) + (M * \omega_t)|_{(x_t, y_t)}$$

where M is the median filtering kernel with size in 3×3 pixels.

Trajectories are formed by concatenating points from subsequent frames: P_t, P_{t+1}, P_{t+2}. To restrain tracking drifting, the length of the trajectories is limited to L frames. Static trajectories and trajectories with sudden large displacements are filtered out in post processing steps. To ensure dense coverage of trajectories, a new point is added if no tracked point is found in a $W \times W$ neighborhood.

3.2. Descriptors

To depict motion pattern, descriptors are computed within a space–time volume aligned with the dense trajectories. The volume size is $N \times N \times L$, where N is in pixels and L is the frame length. Considering the structure information, the volume is further divided into a smaller size $n_\sigma \times n_\sigma \times n_\tau$.

Three types of descriptors, namely HoG (Histograms of Oriented Gradients) [60], HoF (Histogram of Optical Flow) [26] and MBH (Motion Boundary Histograms) [61] are tested in our system. Each has its own characteristics. HoG is designed to encode static appearance information while HoF is good at capturing the local motion. Optical flow represents the absolute between continuous frames. By computing the gradient of the optical flow, MBH keeps the relative motions between pixels and removes constant motions from camera.

3.3. Codebook generation and vector quantization

A large amount of features (in the order of 10^5 or even 10^6 in our experiments) will be extracted along the dense trajectories, which poses big difficulties for learning and classification. The core strength of the Bag-of-Features approach is to form a histogram like compact representation for each candidate by mapping computed features to a codebook. Usually codebook is generated by clustering on training features. Cluster centers are the vocabulary words. Then features are assigned to its nearest words. A histogram like sparse vector is finally computed by the occurrence counts of words.

3.4. Learning action patterns using support vector machine

To learn and predict worker actions, a non-linear support vector machine with $RBF - \chi^2$ kernel is adopted. Various descriptors can be combined using the following approach [26]:

$$K(H_i, H_j) = \exp\left(-\sum_{c \in C} \frac{1}{A_c} D_c(H_i, H_j)\right)$$

where $D_c(H_i, H_j) = \frac{1}{2}\sum_{n=1}^{v} \frac{(h_{in} - h_{jn})^2}{h_{in} + h_{jn}}$. v is the vocabulary size. A_c is the mean value of the distances between all training samples for a channel c. For multi-class classification, SVM classifier is trained using a one-against-rest strategy for each action type. During testing, classifier with the highest confidential value will dominate the action type.

4. Dataset

Facing challenges discussed in Section 2, a new worker action dataset is established. Videos were recorded with handheld camcorders from four construction sites. The ongoing projects in these sites are mainly reinforced concrete buildings. We focus on direct work according to CII's definition [7]. At current stage, the scope is limited to manual work with handheld tools. Therefore workers working with heavy machines or vehicles are out of the range. After two months observation, 11 frequently observed actions are selected to form the dataset. They are "LayBrick", "Transporting", "CutPlate", "Drilling", "TieRebar", "Nailing", "Plastering", "Shoveling", "Bolting", "Welding", "Sawing". A wide range of trades, namely carpenter, ironworker, mason, plasterer and welder, are covered.

While capturing, different weathers, illuminations, points of views, scales and occlusions are covered. Fig. 2 shows examples of various view angles in nailing and tying rebar. The gender of construction workers, as well as their skill levels, is also considered. Most importantly, the recorded worker actions were completely unscripted, unrehearsed and undirected.

Finally, the recorded videos are segmented into 1176 clips with action type annotated manually. Each clip only contains a single action type. Snapshots of video frames from different actions are shown in Fig. 3. Notice that in the 11 types of actions, some of them are cyclic with clear starting point and ending point, such as "LayBrick", "CutPlate","Drilling", "TieRebar","Nailing", "Shoveling" and "Bolting". They are segmented by the action cycle, such as laying a brick, drilling a hole or tying a knot. Some of the actions are repetitive with action primitives, such as "Transporting", "Plas tering","Sawing", and "Welding". It is either difficult to define a clear beginning and ending point for these types of actions. Or even with a task duration, it is too long to be an independent unit for action classification. For example, plastering a wall can take tens of minutes. Cutting a wooden plate into two pieces by sawing may need at least several minutes. For these types of actions, we segment them by action primitives. Usually a video clip contains 8–10 repeated primitives in several seconds duration.

Some features of the proposed dataset are as follows. First, compared to other human action datasets, which are either coarse grained [14,17] or fine grained [20–22], our dataset is combined with both fine grained and coarse grained actions. Some construction tasks involve both limb movement and figure movement. Second, worker actions exhibit low inter-class variability and high intra-class variability. Third, since the recorded videos are completely unscripted and undirected, unexpected situation happens

Fig. 2. Examples of various view angles.

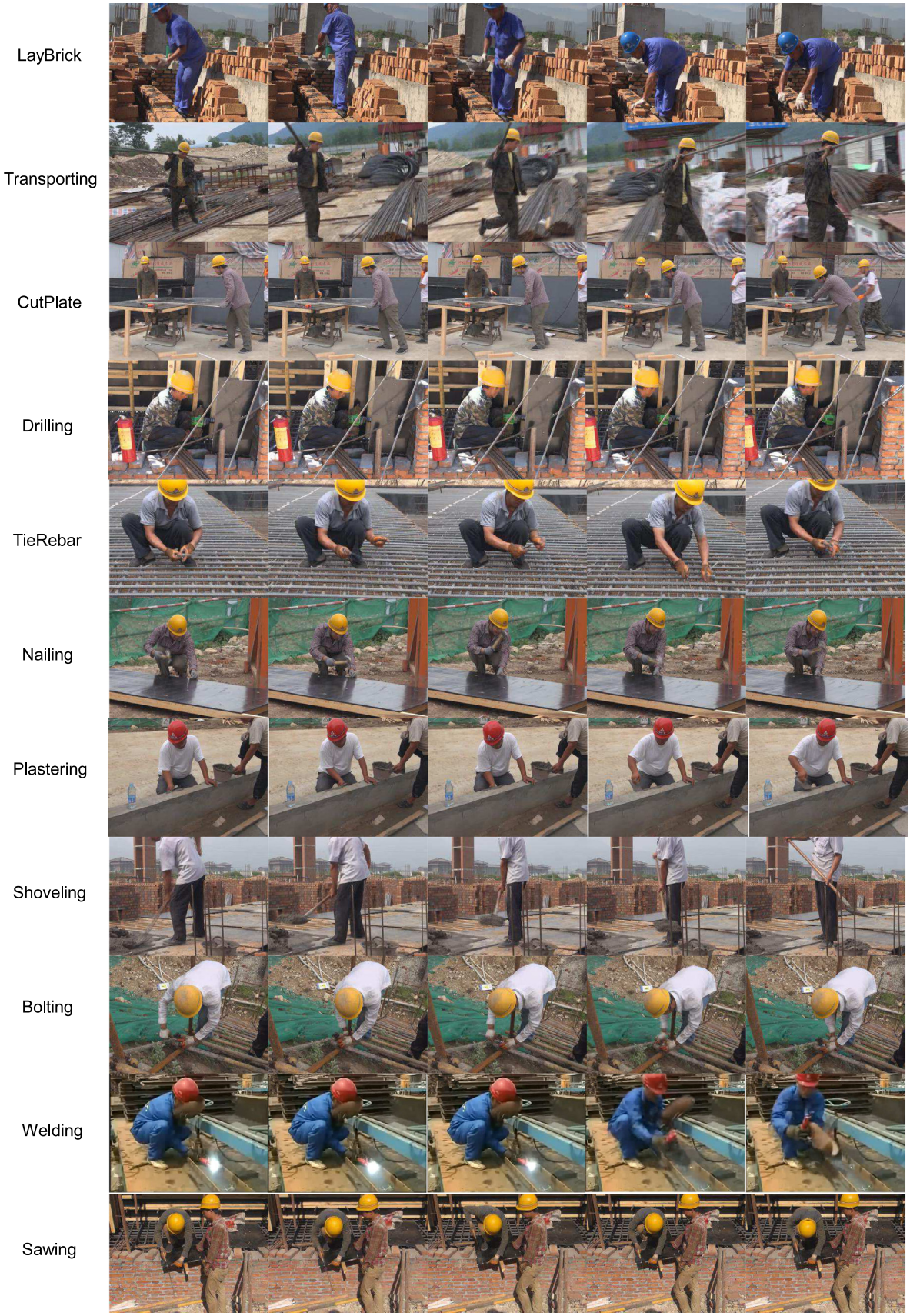

Fig. 3. Snapshots of all actions in our dataset.

Table 1
Statistical information of the proposed dataset.

Action type	Number of Clips	Number of workers	Mean of clips length (s)	Variance of clips length (s)
LayBrick	190	18	9.8	3.6
Transporting	54	25	5.1	1.7
CutPlate	53	7	13.1	5.6
Drilling	58	5	5.9	2.4
TieRebar	157	10	5.1	1.4
Nailing	132	17	5.7	2.5
Plastering	168	12	9.8	5.2
Shoveling	185	22	3.8	1.7
Bolting	79	18	7.2	2.8
Welding	50	4	3.6	2.0
Sawing	50	8	5.9	1.1
Average	107	13	6.8	2.7

Table 2
Comparison with other datasets.

Reference	Dataset	Number of clips	Number of action types	Length of video clips (s)
Gong et al. [43]	Backhoe	150	3	10
	Worker	300	5	5
Golparvar-Fard et al. [44]	Excavator	627	3	6
	Truck	233	3	9
The proposed dataset	Worker	1176	11	6.8

occasionally. For example, in the videos there are interrupted action cycle. We still keep these data since that is what will also happen in real applications. Fourth, though we mainly focus on construction tasks performed by a single worker, we reserve those video clips containing small cooperations by multiple workers. As shown in Fig. 3, two workers are cooperating to cut a plate. Sometimes it is common that a second worker just happens to be in the camera view and introduces some disturbing movement. In a word, we tried to keep all the real situations in the proposed dataset.

All video are in the resolution of 320 by 240 with the frame rate of 30 f/s. The statistical information of the proposed data set is summarized in Table 1. It can be seen that the average number of clips per each action type is 107 performed by 13 workers. The average video length for all action types is 6.8 s with a 2.7 s variation. The variance per each action type is relatively high due to aforementioned data features.

As can be seen in Table 2, compared to the datasets from our two closely related references [43,44], the proposed dataset has the most video clips covering the biggest number of action categories.

5. Experimental results

In the experiment, dense trajectories are computed from inputting videos using Wang's method [34]. Parameters are kept the same as in [34], $W = 5$, $L = 15$, $N = 32$, $n_\sigma = 2$, $n_\tau = 3$. Three types of descriptors HoG, HoF and MBH are computed along the trajectories to depict the motion. The descriptor size is 96 for HoG with 8 bins quantization and 108 for HoF with 9 bins quantization (with a zero bin). Specifically, the MBH descriptor is split into horizontal component MBHx and vertical components MBHy, whose size are both 96.

The proposed data set is divided into four independent groups, where videos in separate groups are taken from different actors. We apply Leave-One-Group-Out Cross-Validation for training and testing. During training, a subset of 100,000 randomly selected features is clustered using k-means to generate codebooks for each description channel. Then all features are quantized by assigning to the nearest vocabulary word using Euclidean distance. SVM classifiers with RBF kernel are trained using quadratic programming. All experimental results given below are averaged through four folds cross validation.

To evaluate the performance results, the confusion matrix and average per class accuracy have been adopted. The confusion matrix $C(i,j)$ is a percent count of observations known to be type i but predicted as type j. Each column of the matrix represents the instances in a predicted action class, while each row represents the instances in an actual action class. The average per class accuracy is defined as:

$$ACC = \sum_{i=j=1}^{i=j=N} C(i,j)/N$$

where N is the number of action categories.

The performance of the system has been tested on each individual descriptor (HoG, HoF, MBH), and also the combination of all descriptors using the multi-channel approach as aforementioned. Notice the performance of MBH is the combination of MBHx and MBHy using multi-channel approach. The impact of various codebook size on the system has been investigated as well. Lastly, we performed the algorithm from Gong et al. [43] on the proposed dataset for comparison.

5.1. Experimental results on various descriptors

Though it has been reported [34] that the combination of all descriptors outperforms each individual descriptor on action recognition, the result is not achieved on coarse-grained and fine-grained mixed action dataset. It would still be safe to test on individual descriptor first and then their combination. To test the impact of various descriptors on system performance, the number of visual words per descriptor is fixed to 4000, which is shown to perform well on a wide range of datasets [34]. As can been seen in Fig. 4, the average confusion matrices for each type of descriptor and their combination are given separately. Generally speaking, the top three action categories with high accuracy are "LayBrick", "TieRebar" and "Transporting". And the bottom three actions are "Drilling", "Bolting" and "Sawing". One reason is that the former three categories contain obvious movement and relatively standard workflow. For example, a common "LayBrick" flow is: "pick up a brick", "get mortar with a trowel", "smear the mortar", "place the brick", "knock the brick with the trowel to fasten". The latter three categories do not have either large movement or consistent workflow. For example, when drilling, the worker's body nearly holds still, only with the bit spinning rapidly, which is really difficult to capture in a 30 f/s video. For these types of actions, supplementary information such as tools recognition might be helpful to enhance the performance [62]. Notice that dense trajectories are formed by densely sampling all over the images, which is to say, sematic background information is also critical for action prediction. Future study may seek to encode background and foreground separately and assign the corresponding features different weights according to the magnitude of motions. For example, for those action types with large motions, features on foreground area should weigh more than those on background area and vice versa.

For a clearer comparison, we plot the average accuracy per each action type per descriptor together in Fig. 5, where red[1] line with asterisk markers represents descriptor HOG, green line with circle markers is for HoF, blue line with plus sign markers is for MBH,

[1] For interpretation of color in Figs. 5 and 7, the reader is referred to the web version of this article.

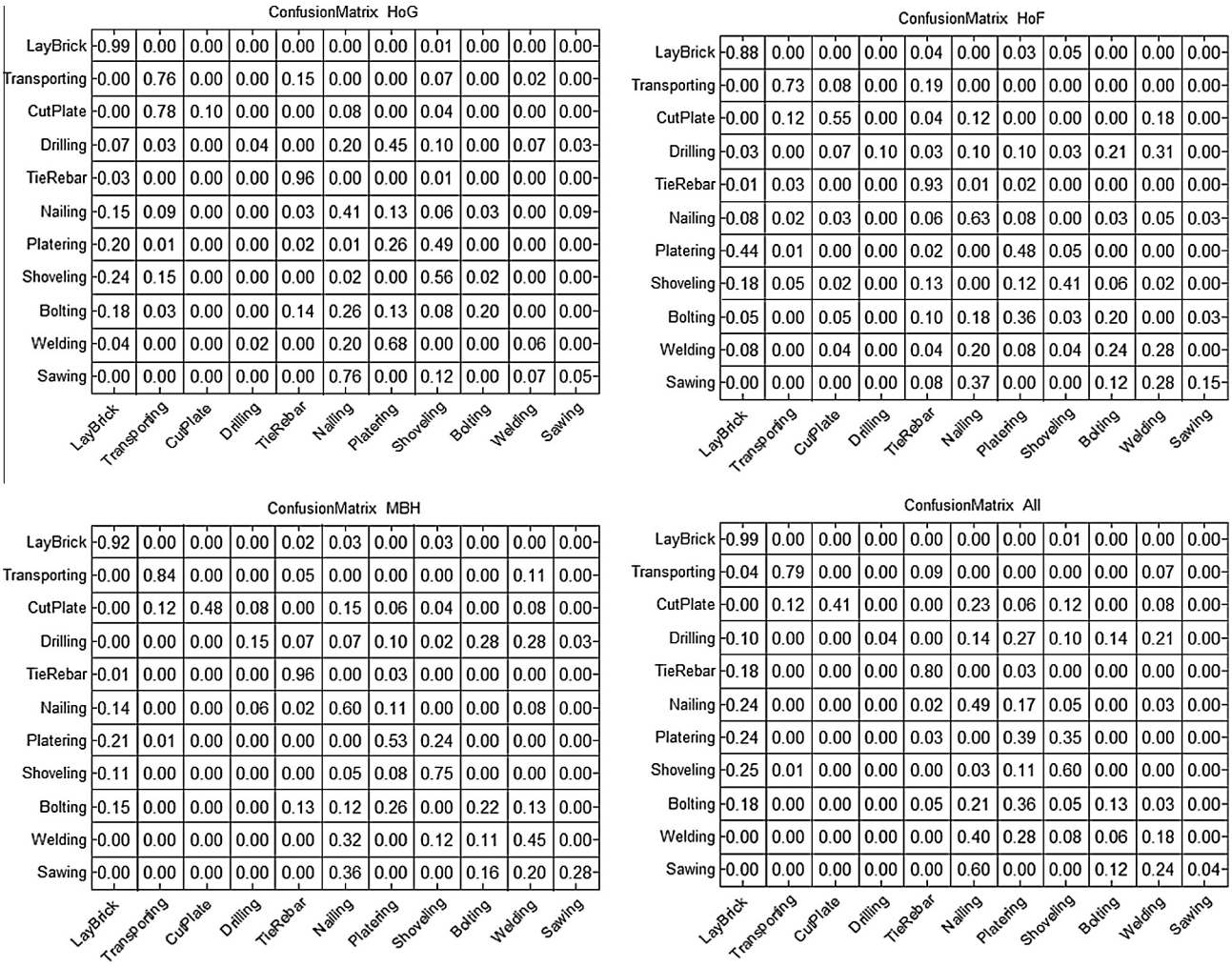

Fig. 4. Confusion matrix of action recognition with various descriptors.

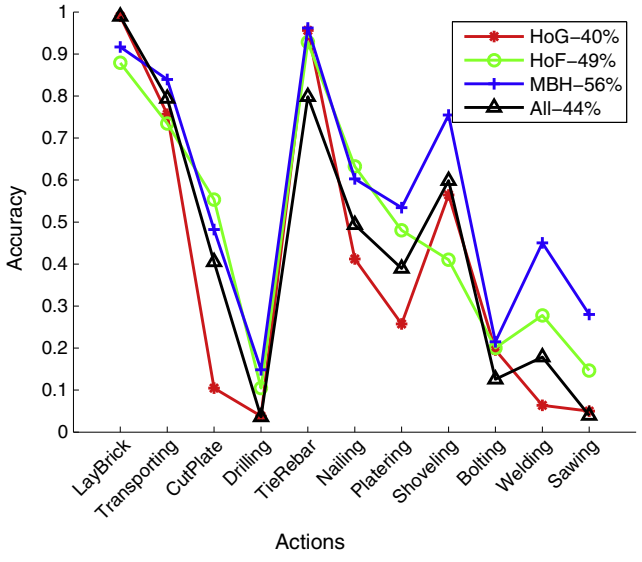

Fig. 5. Recognition accuracy per action type with various descriptors.

and black line with triangle markers represents the combination of all three descriptors. The average per class accuracy for each descriptor is 40%, 49%, 56% and 44% for descriptor HoG, HoF, MBH and the combination of them separately. It can be seen that except for action type "LayBrick" and "CutPlate", where MBH exhibits slightly weaker accuracy compared to HoG and HoF, MBH achieves the highest per action accuracy elsewhere and gives out the best overall performance of 56%. This is mainly due to its ability to suppress camera motion and capture local motion better. Unexpectedly, the combination of all descriptors does not perform as well as in [34], where it achieved the best performance in all tested nine datasets. One possible explanation is that actions in our dataset are not as consistent as those in other datasets. Some actions are coarse-grained, such as "Transporting", "CutPlate" and "Shoveling". Some are fine-grained, such as "TieRebar" and "Bolting". A few categories are somewhere in between, such as "LayBrick" and "Nailing", where both coarse body movement and fine hand movement are involved. A naive combination of all descriptors may affect the discrimination ability adversely.

5.2. Experimental results on different codebook size

Codebook generation is a key step in Bag-of-Feature pipeline. The size of codebook dominates the granularity of motion description and then affects the final action recognition performance. Furthermore, since feature quantization is one of the most computationally expensive step, the codebook size is closely related with the computation complexity. Previously, a default value of 4000 is adopted. However, the most appropriate codebook size is unexplored. To investigate the impact of the codebook size

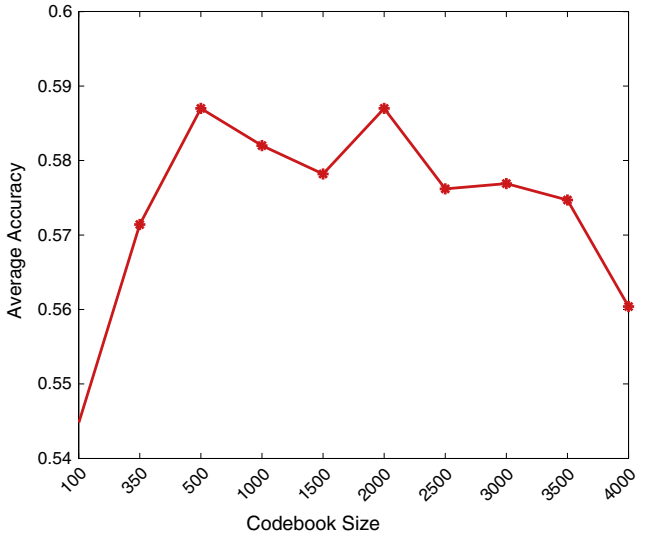

Fig. 6. The average accuracy under different codebook size.

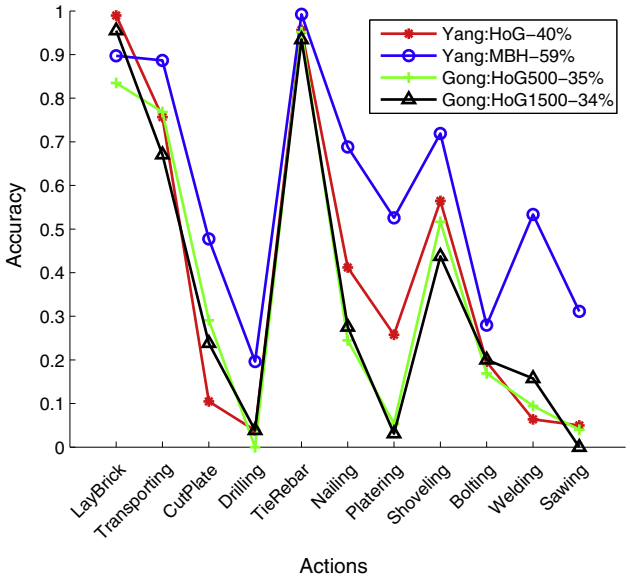

Fig. 7. Comparison to Gong's method on the proposed dataset.

on the system performance, the codebook size has been tuned from 100 to 4000 with average accuracy of action recognition shown in Fig. 6. At this time, the descriptor with best performance – MBH is chosen. Obviously there are two equal peaks at size 500 and 2000. And the default value of 4000 only performs better than the worst case at size 100. Considering the computation complexity, it can be suggested that 500 is the most appropriate codebook size. For reference, the best codebook size in [43,44] are 1500 and 350 separately.

5.3. Computation time

Time used for action learning and classification is consumed in the steps of feature computation, codebook formation, vector quantization, classifiers training and action type prediction. Except feature computation which runs in C++ language, other stages are all implemented in MATLAB. The run-time is obtained on a desktop PC with a 3.5 GHz quadcore Intel CPU and 16 GB RAM. Keeping parameters set as aforementioned, the time for computing all three types of descriptors is around 2 frames/s. Time for codebook formation and vector quantization depends largely on the selected codebook size. With codebook size 4000, average time for codebook formation on 100,000 features is around 1000 s. Vector quantization takes approximately 85 s per each video clip. Time for training 11 classifiers and predicting action type was benchmarked as 514 and 174 s separately. Notice codebook formation and classifiers training can all be done as offline processes. So the main computation bottleneck is at the vector quantization stage.

5.4. Comparison to the state-of-the-art results

As indicated in Table 2, there are two closely related references. Gong et al. [43] tested their method both on equipment and worker dataset, while Golparvar-Fard et al. [44] mainly focused on earthmoving equipment actions. Since we only discuss worker action recognition in this paper, Gong's method is applied on our dataset for comparison. Specifically, Gong et al. [43] adopted 3D Harris corner detector [28] to model image sequences with HoG or HoF descriptors. And it was claimed that the configuration of 1500 code words, HoG descriptor, and naive Bayesian model produces the best classification results [43]. Since we use discriminative model (SVM) other than generative model (Bayesian model) in our system, the comparison is limited to the video modeling stage. We applied 3D Harris corner detector and HoG to our dataset with SVM for recognition. All parameters are set the same as they are in [43]. As shown in Fig. 7, red line with asterisk markers is our method with codebook size 500 and HoG descriptor, blue line with circle markers is our method with codebook size 500 and MBH descriptor, green line with plus sign markers is Gong's method with codebook size 500 and HoG descriptor, and black line with triangle markers is Gong's method with codebook size 1500 and HoG descriptor. The average accuracy for these four setups are 40%, 59%, 35% and 34% separately. It can be concluded that our method performs better than Gong's method. With the same type of descriptor, dense trajectories models motion better than 3D Harris corner. And the MBH descriptor enhances the system performance dramatically.

6. Conclusions

In this paper, we studied vision-based worker action recognition using the Bag-of-Feature framework. A cutting-edge video representation method – dense trajectories was adopted. Three types of descriptors, namely HoG, HoF and MBH, and their combination were tested for performance evaluation. The multi-class SVM with non-linear RBF kernel was applied for training and classification. A new real world dataset were established for system validation with totally 1176 video clips, covering 11 categories of common worker actions. Several challenging situations, such as view angle change, illumination change, interrupted workflow, and interaction between multiple workers, are involved in the proposed dataset. Experimental results showed that the system achieved the best performance with average accuracy of 59% under the configuration of MBH descriptor and codebook size 500, outperforming Gong et al.'s method [43] by 24%. The system holds a promising potential for future real world application.

However, several limitations exist and may be improved in the future. From the algorithm aspect, recently convolutional neural network (CNN) has exhibit good potentials in action recognition [63,64], especially on super large scale dataset (with millions of samples). Considering it will be difficult to achieve such a scale in construction dataset, it is still unknown how CNN will perform on construction scenario. A comparison on both accuracy and computational cost is needed between CNN and feature based method.

From the construction aspect, though we have established a worker action dataset, a comprehensive taxonomy defining worker activity lexicon and hierarchy [65] is still missing. Additionally, each activity under the taxonomy should be modeled statistically as temporally structured actions [22]. Only in this way can automated activity analysis in long videos be realized. Lastly, in this paper, since dense sampling captures information not only from workers but also from background scenario, we did not take tools into account explicitly. However, it has been shown that combining tool detection can enhance action recognition [62]. In a very recent study, Jain et al. [66] encoded 15,000 object categories for actions without explicit detection and proved that adding object encodings improved action classification and localization. We plan to explore tool encodings with worker actions in the future.

Acknowledgement

The work is supported by National Natural Science Foundation of China Nos. 51208425 and 51278420.

References

[1] P. Teicholz, Labor-productivity declines in the construction industry: causes and remedies (another look), AECbytes Viewpoint 67 (2013).
[2] M.C. Gouett, C.T. Haas, P.M. Goodrum, C.H. Caldas, Activity analysis for direct-work rate improvement in construction, J. Construct. Eng. Manage. 137 (2011) 1117–1124.
[3] A. Khosrowpour, J.C. Niebles, M. Golparvar-Fard, Vision-based workface assessment using depth images for activity analysis of interior construction operations, Autom. Construct. 48 (2014) 74–87.
[4] T. Cheng, J. Teizer, G.C. Migliaccio, U.C. Gatti, Automated task-level activity analysis through fusion of real time location sensors and worker's thoracic posture data, Autom. Construct. 29 (2013) 24–39.
[5] M.-W. Park, C. Koch, I. Brilakis, Three-dimensional tracking of construction resources using an on-site camera system, J. Comput. Civil Eng. 26 (2012) 541–549.
[6] E. Rezazadeh Azar, S. Dickinson, B. McCabe, Server–customer interaction tracker: computer vision–based system to estimate dirt-loading cycles, J. Construct. Eng. Manage. 139 (2012) 785–794.
[7] Construction Industry Institute (CII) (Ed.), IR252.2a – Guide to Activity Analysis, Construction Industry Institute, Austin, TX, USA, 2010.
[8] M.E. Shehata, K.M. El-Gohary, Towards improving construction labor productivity and projects performance, Alex. Eng. J. 50 (2011) 321–330.
[9] J. Teizer, Status quo and open challenges in vision-based sensing and tracking of temporary resources on infrastructure construction sites, Adv. Eng. Inf. 29 (2015) 225–238.
[10] J. Seo, S. Han, S. Lee, H. Kim, Computer vision techniques for construction safety and health monitoring, Adv. Eng. Inf. 29 (2015) 239–251.
[11] J. Yang, M.-W. Park, P.A. Vela, M. Golparvar-Fard, Construction performance monitoring via still images, time-lapse photos, and video streams: now, tomorrow, and the future, Adv. Eng. Inf. 29 (2015) 211–224.
[12] T.B. Moeslund, A. Hilton, V. Krüger, A survey of advances in vision-based human motion capture and analysis, Comp. Vis. Image Understand. 104 (2006) 90–126.
[13] M. Blank, L. Gorelick, E. Shechtman, M. Irani, R. Basri, Actions as space–time shapes, Tenth IEEE International Conference on Computer Vision, 2005, ICCV 2005, vol. 2, IEEE, 2005, pp. 1395–1402.
[14] C. Schüldt, I. Laptev, B. Caputo, Recognizing human actions: a local svm approach, Proceedings of the 17th International Conference on Pattern Recognition, 2004, ICPR 2004, vol. 3, IEEE, 2004, pp. 32–36.
[15] D. Tran, A. Sorokin, Human activity recognition with metric learning, in: Computer Vision–ECCV 2008, Springer, 2008, pp. 548–561.
[16] M.D. Rodriguez, J. Ahmed, M. Shah, Action mach a spatio-temporal maximum average correlation height filter for action recognition, in: IEEE Conference on Computer Vision and Pattern Recognition, 2008, CVPR 2008, IEEE, 2008, pp. 1–8.
[17] J.C. Niebles, C.-W. Chen, L. Fei-Fei, Modeling temporal structure of decomposable motion segments for activity classification, in: Computer Vision–ECCV 2010, Springer, 2010, pp. 392–405.
[18] F. De La Torre, J. Hodgins, A. Bargteil, X. Martin, J. Macey, A. Collado, P. Beltran, Guide to the Carnegie Mellon University multimodal activity (CMU-MMAC) database, Technical Report CMU-RI-TR-08-22, Carnegie Mellon University, 2008.
[19] M. Tenorth, J. Bandouch, M. Beetz, The tum kitchen data set of everyday manipulation activities for motion tracking and action recognition, in: IEEE 12th International Conference on Computer Vision Workshops (ICCV Workshops), 2009, IEEE, 2009, pp. 1089–1096.
[20] M. Rohrbach, S. Amin, M. Andriluka, B. Schiele, A database for fine grained activity detection of cooking activities, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2012, IEEE, 2012, pp. 1194–1201.
[21] S. Stein, S.J. McKenna, Combining embedded accelerometers with computer vision for recognizing food preparation activities, in: Proceedings of the 2013 ACM International Joint Conference on Pervasive and Ubiquitous Computing, ACM, 2013, pp. 729–738.
[22] H. Kuehne, A. Arslan, T. Serre, The language of actions: recovering the syntax and semantics of goal-directed human activities, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, IEEE, 2014, pp. 780–787.
[23] J. Yang, Z. Shi, Z. Wu, Automatic recognition of construction worker activities using dense trajectories, in: 32nd International Symposium on Automation and Robotics in Construction and Mining, 2015, pp. 75–81.
[24] D. Weinland, R. Ronfard, E. Boyer, A survey of vision-based methods for action representation, segmentation and recognition, Comp. Vis. Image Understand. 115 (2011) 224–241.
[25] R. Poppe, A survey on vision-based human action recognition, Image Vis. Comput. 28 (2010) 976–990.
[26] I. Laptev, M. Marszałek, C. Schmid, B. Rozenfeld, Learning realistic human actions from movies, in: IEEE Conference on Computer Vision and Pattern Recognition, 2008, CVPR 2008, IEEE, 2008, pp. 1–8.
[27] H. Wang, A. Kläser, C. Schmid, C.-L. Liu, Dense trajectories and motion boundary descriptors for action recognition, Int. J. Comp. Vis. 103 (2013) 60–79.
[28] I. Laptev, On space–time interest points, Int. J. Comp. Vis. 64 (2005) 107–123.
[29] P. Dollár, V. Rabaud, G. Cottrell, S. Belongie, Behavior recognition via sparse spatio-temporal features, in: 2nd Joint IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance, 2005, IEEE, 2005, pp. 65–72.
[30] G. Willems, T. Tuytelaars, L. Van Gool, An efficient dense and scale-invariant spatio-temporal interest point detector, in: Computer Vision–ECCV 2008, Springer, 2008, pp. 650–663.
[31] P. Scovanner, S. Ali, M. Shah, A 3-dimensional sift descriptor and its application to action recognition, in: Proceedings of the 15th International Conference on Multimedia, ACM, 2007, pp. 357–360.
[32] A. Klaser, M. Marszałek, C. Schmid, A spatio-temporal descriptor based on 3d-gradients, in: BMVC 2008-19th British Machine Vision Conference, British Machine Vision Association, 2008, pp. 995–1004.
[33] L. Yeffet, L. Wolf, Local trinary patterns for human action recognition, in: IEEE 12th International Conference on Computer Vision, 2009, IEEE, 2009, pp. 492–497.
[34] H. Wang, A. Kläser, C. Schmid, C.-L. Liu, Action recognition by dense trajectories, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2011, IEEE, 2011, pp. 3169–3176.
[35] J. Zou, H. Kim, Using hue, saturation, and value color space for hydraulic excavator idle time analysis, J. Comput. Civil Eng. 21 (2007) 238–246.
[36] J. Gong, C.H. Caldas, An intelligent video computing method for automated productivity analysis of cyclic construction operations, Comput. Civil Eng. (2009) (2009) 64–73.
[37] J. Gong, C.H. Caldas, Computer vision-based video interpretation model for automated productivity analysis of construction operations, J. Comput. Civil Eng. 24 (3) (2009) 252–263.
[38] J. Yang, P. Vela, J. Teizer, Z. Shi, Vision-based tower crane tracking for understanding construction activity, J. Comput. Civil Eng. 28 (2012) 103–112.
[39] A. Peddi, L. Huan, Y. Bai, S. Kim, Development of human pose analyzing algorithms for the determination of construction productivity in real-time, Construction Research Congress, vol. 1, ASCE, Seattle, WA, 2009, pp. 1–20.
[40] J. Gong, C.H. Caldas, An object recognition, tracking, and contextual reasoning-based video interpretation method for rapid productivity analysis of construction operations, Autom. Construct. 20 (2011) 1211–1226.
[41] M. Bügler, G. Ogunmakin, J. Teizer, P.A. Vela, A. Borrmann, A comprehensive methodology for vision-based progress and activity estimation of excavation processes for productivity assessment, in: Proceedings of the 21st International Workshop: Intelligent Computing in Engineering (EG-ICE), Cardiff, Wales, 2014.
[42] M.-W. Park, Automated 3D Vision-Based Tracking of Construction Entities, Georgia Institute of Technology, Atlanta, GA, USA, 2012.
[43] J. Gong, C.H. Caldas, C. Gordon, Learning and classifying actions of construction workers and equipment using bag-of-video-feature-words and bayesian network models, Adv. Eng. Inf. 25 (2011) 771–782.
[44] M. Golparvar-Fard, A. Heydarian, J.C. Niebles, Vision-based action recognition of earthmoving equipment using spatio-temporal features and support vector machine classifiers, Adv. Eng. Inf. 27 (2013) 652–663.
[45] I.T. Weerasinghe, J.Y. Ruwanpura, J.E. Boyd, A.F. Habib, Application of microsoft kinect sensor for tracking construction workers, in: Construction Research Congress, 2012, pp. 858–867.
[46] Y. Turkan, F. Bosché, C.T. Haas, R. Haas, Toward automated earned value tracking using 3d imaging tools, J. Construct. Eng. Manage. 139 (2012) 423–433.
[47] J.M. Chaquet, E.J. Carmona, A. Fernández-Caballero, A survey of video datasets for human action and activity recognition, Comp. Vis. Image Understand. 117 (2013) 633–659.
[48] J. Liu, J. Luo, M. Shah, Recognizing realistic actions from videos in the wild, in: IEEE Conference on Computer Vision and Pattern Recognition, 2009, CVPR 2009, IEEE, 2009, pp. 1996–2003.

[49] M. Marszalek, I. Laptev, C. Schmid, Actions in context, in: IEEE Conference on Computer Vision and Pattern Recognition, 2009, CVPR 2009, IEEE, 2009, pp. 2929–2936.
[50] K.K. Reddy, M. Shah, Recognizing 50 human action categories of web videos, Mach. Vis. Appl. 24 (2013) 971–981.
[51] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, T. Serre, Hmdb: a large video database for human motion recognition, in: 2011 IEEE International Conference on Computer Vision (ICCV), IEEE, 2011, pp. 2556–2563.
[52] S. Oh, A. Hoogs, A. Perera, N. Cuntoor, C.-C. Chen, J.T. Lee, S. Mukherjee, J. Aggarwal, H. Lee, L. Davis, et al., A large-scale benchmark dataset for event recognition in surveillance video, in: 2011 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2011, pp. 3153–3160.
[53] M.S. Ryoo, J.K. Aggarwal, Spatio-temporal relationship match: video structure comparison for recognition of complex human activities, in: IEEE 12th International Conference on Computer Vision, 2009, IEEE, 2009, pp. 1593–1600.
[54] R. Fisher, J. Santos-Victor, J. Crowley, Caviar: Context aware vision using image-based active recognition, 2005.
[55] J. Ferryman, A. Ellis, Pets2010: dataset and challenge, in: 2010 Seventh IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), IEEE, 2010, pp. 143–150.
[56] R. Fisher, Behave: computer-assisted prescreening of video streams for unusual activities, 2004.
[57] R. Messing, C. Pal, H. Kautz, Activity recognition using the velocity histories of tracked keypoints, in: IEEE 12th International Conference on Computer Vision, 2009, IEEE, 2009, pp. 104–111.
[58] A. Voulodimos, D. Kosmopoulos, G. Vasileiou, E. Sardis, A. Doulamis, V. Anagnostopoulos, C. Lalos, T. Varvarigou, A dataset for workflow recognition in industrial scenes, in: 18th IEEE International Conference on Image Processing (ICIP), 2011, 2011, pp. 3249–3252.
[59] N. Padoy, T. Blum, H. Feussner, M.-O. Berger, N. Navab, On-line recognition of surgical activity for monitoring in the operating room, in: AAAI, 2008, pp. 1718–1724.
[60] N. Dalal, B. Triggs, Histograms of oriented gradients for human detection, IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2005, CVPR 2005, vol. 1, IEEE, 2005, pp. 886–893.
[61] N. Dalal, B. Triggs, C. Schmid, Human detection using oriented histograms of flow and appearance, in: Computer Vision–ECCV 2006, Springer, 2006, pp. 428–441.
[62] J.Y. Kim, C.H. Caldas, Vision-based action recognition in the internal construction site using interactions between worker actions and construction objects, in: International Symposium on Automation and Robotics in Construction and Mining, 2013, pp. 661–668.
[63] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, L. Fei-Fei, Large-scale video classification with convolutional neural networks, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, IEEE, 2014, pp. 1725–1732.
[64] Y. Du, W. Wang, L. Wang, Hierarchical recurrent neural network for skeleton based action recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 1110–1118.
[65] F. Caba Heilbron, V. Escorcia, B. Ghanem, J. Carlos Niebles, Activitynet: a large-scale video benchmark for human activity understanding, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 961–970.
[66] M. Jain, J.C. van Gemert, C.G. Snoek, What do 15,000 object categories tell us about classifying and localizing actions? in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 46–55.